AUTOMATED RECOGNITION OF ADVANCED VIBRATION FEATURES FOR MACHINERY FAULT CLASSIFICATION

Katherine McClintic, Robert Campbell, Gregory Babich,
Amulya Garga, Jeffery Banks, Michael Thurston, Carl Byington

Applied Research Laboratory
The Pennsylvania State University
P.O. Box 30
State College, PA 16804-0030

Abstract: Advanced condition monitoring systems use pattern recognition and automated reasoning on features extracted from sensor data to assess the current health of a component. This paper will evaluate pattern recognition techniques for classifying the "stage of fault" using transitional failure data for commercial grade gearboxes. Features will be extracted from accelerometer data obtained on the Mechanical Diagnostic Testbed (MDTB) at Penn State Applied Research Lab. The ARL CBM Features toolbox, a MATLAB-based toolbox containing most of the traditional HUMS features and several novel features, will be used to perform feature extraction. Several classifiers and training methods will be evaluated, as well as the effect of using different dimension-reduction techniques on classification. The results obtained using the transitional failure data sets will contribute to enhanced health monitoring techniques and improved machinery health prognostic estimates.

Keywords: Classification; gearbox tooth breakage; MDTB; pattern recognition

Introduction: Penn State University Applied Research Laboratory (ARL) is contributing to the diagnostics and prognostics development for aircraft systems using statistical pattern recognition and sensor fusion. Analysis was conducted using a software package developed by ARL called the Shell Enhanced Pattern Recognition Advanced Toolbox (SEPARAT). For this analysis, features were extracted from gearbox run-to-failure accelerometer data acquired on the Mechanical Diagnostics Test Bed (MDTB) at ARL. Based upon borescope ground truth, the data was segmented into three classes: no failure, 1-2 teeth broken, and 2-8 teeth broken. Various classifiers, dimensionality reduction techniques, and training methods were evaluated for their ability to classify stage of fault.

Feature Extraction: In principle, information concerning the relative condition of the monitored machine can be extracted from the vibration signature, and inferences can be made about the health by comparing the vibration signal with previous signals to identify any anomalous conditions that may be occurring. In practice, however, such direct comparisons are not effective mainly due to the large variations between subsequent

signals. Instead, several more useful techniques have been developed over the years that involve feature extraction from the vibration signature [9]. Generally these features are more stable and well behaved than the raw signature data itself. In addition, the features constitute a reduced data set, because one feature value may represent an entire snapshot of data, thus facilitating additional analysis such as pattern recognition for diagnostics and feature tracking for prognostics. Moreover, the use of feature values instead of raw vibration data will become extremely important as wireless applications, with greater bandwidth restrictions, become more widely used.

The feature extraction method may require several steps, depending on the type of feature being calculated. Some features are calculated using the "conditioned" raw signal, while others use a time-synchronous averaged signal that has been filtered to remove the "common" spectral components. ARL developed a CBM Features Toolbox that allows these features to be calculated systematically.

Pattern Recognition Overview: The classification techniques used for this analysis are included in SEPARAT as well as neural networks, Gaussian classifiers, statistical analysis and feature reduction techniques. A discussion of some of the key pattern recognition terminology is provided below.

Feature extraction, as discussed above, is the process of reducing measured signals into feature vectors. Classifier design, also called training, is the process of determining feature space partitions so that unlabeled vectors can be given a class label. Evaluation is the process of testing the design of both the classifier and its inputs. If the evaluation is unsatisfactory, other classifier structures, features and/or attributes, must be sought; otherwise, a satisfactory evaluation indicates the selected attributes, features, and classifier can be incorporated into the application.

Optimal Classification: The goal of pattern classification is to assign a physical object or process to one of c pre-defined classes [1]. The idealized Bayes decision strategy yields a classifier that is optimal (i.e., the classification error rate is minimal). This concept is paramount to pattern classification regardless of the particular technique used. Let $\mathbf{x}$ be a random variable with d-components (features) which obeys the class conditional probability density function $p(\mathbf{x}|\omega_i)$, where ω_i represents one of c possible objects or processes that are of interest, and $P(\omega_i)$ represents the a priori probability that ω_i occurs. The state-conditional a posteriori probability can be expressed by Bayes rule [1]:

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})}, \tag{1}$$

where

$$p(\mathbf{x}) = \sum_{i=1}^{c} p(\mathbf{x}|\omega_i)P(\omega_i). \tag{2}$$

The optimal decision rule with the smallest possible error is given as [1]:

Decide that $\mathbf{x}$ belongs to class ω_i iff $P(\omega_i|\mathbf{x}) > P(\omega_j|\mathbf{x})$ for all $j \neq i$. (3)

An equivalent decision rule is given by:

Decide that $\mathbf{x}$ belongs to class ω_i iff $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $j \neq i$ (4)

where the discriminants, $g_i(\mathbf{x})$, are defined in the present context of the optimal classifier as

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})}. \tag{5}$$

The decision boundaries between the classes labeled ω_i and ω_j consist of the points in feature space where $g_i(\mathbf{x}) = g_j(\mathbf{x})$. Decision boundaries partition d-dimensional feature space into the decision regions that are used to classify unlabeled feature vectors.

Discriminant Functions: Because Eq. (4) compares all discriminant function outputs to find the maximum, only the relative values of the outputs are important. Therefore, equivalent changes can be made to each discriminant function without affecting the classification results. In other words, the decision boundaries are not changed. As long as each discriminant function is changed in the same way, the classification results will not be changed. Thus Eq. (5) can be simplified by removing the scaling constant $p(\mathbf{x})$ while giving the same classification results [1]:

$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i). \tag{6}$$

The Bayes decision rule defines the lowest possible error rate for a given problem. However, the ideal Bayes approach is not truly practical because it requires a priori knowledge of the distribution and its parameters for each class, which are rarely, if ever, known [1]. In practice, one must either assume class-conditional density models and estimate their parameters or estimate the probability densities, either explicitly or implicitly, from observed data. There are several well-known statistical techniques available. In general, they can be grouped as parametric or nonparametric. Parametric approaches assume that the functional form of the class-conditional density functions, which are described by some parameters, are known. Nonparametric approaches do not assume anything about class-conditional distributions.

Nonparametric Approaches: Some approaches, such as the minimum-distance classifier, have widely been used because they offer intuitive appeal and computational simplicity. Other nonparametric approaches, such as the minimum-squared-error algorithm, use the data to optimize a family of linear discriminant functions. When using nonparametric classifiers, class labels are traditionally assigned based on formula (4).

Minimum distance classifiers are widely referenced throughout the literature [1,2,6,7]. With this type of classifier, unknown feature vectors are assigned the class membership of the nearest sample mean. The discriminant function can be written as

$$\begin{aligned} g_i(\mathbf{x}) &= -(\mathbf{x}-\mathbf{m}_i)^t G_i^{-1}(\mathbf{x}-\mathbf{m}_i) \\ &= -\mathbf{x}^t G_i^{-1}\mathbf{x} + 2\mathbf{x}^t G_i^{-1}\mathbf{x} - \mathbf{m}_i^t G_i^{-1}\mathbf{m}_i \end{aligned} \tag{7}$$

where G_i is a positive definite symmetric weighting matrix. Often the sample covariance matrices are used for G_i; the resulting classifier is quadratic. Another variation of the minimum distance classifier involves using the same weighting matrix is used for each class (i.e., $G_i = G$), which results in a linear discriminant function. Using a common

weighting matrix in Eq. (7), multiplying by ½, and dropping the common quadratic term, the linear minimum-distance discriminant function is obtained:

$$g_i(\mathbf{x}) = \mathbf{x}^t G^{-1}\mathbf{m}_i - \frac{1}{2}\mathbf{m}_i^t G^{-1}\mathbf{m}_i . \tag{8}$$

Two commonly used weighting matrices are the identity matrix and the pooled covariance matrix, which results in the Euclidean and Fisher minimum-distance classifiers, respectively [1,6,7]. Minimum-distance classifiers are trivial to train and implement. Training only requires calculation of the sample means and weighting matrices. Classification only requires calculation of Eq. (7) or Eq. (8) for $i = 1, 2, \ldots, c$, followed by comparisons of the discriminant values (Eq. 4).

The linear classifier is particularly attractive because of its computational simplicity during classification. In some cases, linear discriminant functions arise naturally due to the distribution of the data. The minimum-distance classifier is linear when common weighting matrices are used. The following paragraphs discuss a case were the structure is assumed linear and the weights are found based on that assumption. The general form of the linear discriminant function is given by

$$g_i(\mathbf{x}) = \mathbf{x}^t\mathbf{w}_i + w_{i0}, \tag{9}$$

where $\mathbf{w}_i$ and w_{i0} are the weight vector and bias term for the ith class respectively. A family of linear discriminants can be written as

$$\begin{aligned}\mathbf{g}(\mathbf{x}) &= \begin{bmatrix} g_1(\mathbf{x}) & g_2(\mathbf{x}) & \cdots & g_c(\mathbf{x}) \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{x}^t & 1 \end{bmatrix} \begin{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ w_{10} \end{bmatrix} & \begin{bmatrix} \mathbf{w}_2 \\ w_{20} \end{bmatrix} & \cdots & \begin{bmatrix} \mathbf{w}_{c1} \\ w_{c0} \end{bmatrix} \end{bmatrix}, \\ &= \mathbf{y}W \end{aligned} \tag{10}$$

where $\mathbf{y}$ is the 1-by-(d+1) augmented feature vector and W is the (d+1)-by-c weight matrix. One method of training the discriminants is to solve the matrix equation

$$YW = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_c \end{bmatrix} W = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_c \end{bmatrix} = B, \tag{11}$$

where Y is the n_i-by-(d+1) matrix of augmented training samples for the ith class, and B is the corresponding n_i-by-c target matrix where the ith column contains all ones and the other elements are zeros [1]. Equation (11) can be solved using the pseudo-inverse of Y [1,4]:

$$W = Y^{+}B. \tag{12}$$

This approach minimizes the *trace* of squared error matrix $(YW\text{-}B)^t(YW\text{-}B)$; the resulting discriminant functions used in conjunction with (Eq. 4) are called the minimum-squared-error classifiers.

Error Rate Estimation: Choosing which classifier to use for a specific problem is often difficult. To aid in the selection of an appropriate classifier, rigorous analyses can be made to compare the performance of competing designs. In general, error estimation is

accomplished by designing a classifier on training data, labeling test data, and counting the number of errors (mislabeled samples) to estimate the error rate e. Given that $n(\omega_i)$ is the number of samples from the class ω_i incorrectly labeled by the classifier, a typical error estimate is given by

$$e = \sum_{i=1}^{c} \frac{n(\omega_i)}{n_i} P(\omega_i) \tag{13}$$

Several methods can be used to segment available data into training and test sets. When using the resubstitution procedure, the classifier is trained and tested using the same data. This results in an optimistically biased error rate [1,2,6,7]. The resubstitution error rate estimator is good for finding a lower bound on the Bayes error rate [2]. The available data can be split into two mutually exclusive sets for training and testing. This is known as the holdout procedure, which results in an unbiased estimate of the error rate [2]. However, using the holdout procedure requires the data to be segmented either manually or by some clustering algorithm [2]. In many situations, collecting data is very expensive which results in small data sets. One technique that is useful for small data sets is the leave-one-out procedure [1,2,6,7]. In this procedure, all but one of the available training samples is used to train a classifier. The classifier is then tested with the sample that was left out. This process is repeated until all of the available training samples have had their turn as a test sample. The leave-one-out error estimator is nearly unbiased but has a large variance [6,7].

Experimental Facilities: Without a recorded progression to failure, the ability to perform prognostics, the capability to predict remaining useful life, is nearly impossible. This need for high-fidelity transitional data was identified several years ago at ARL. Since then several test beds have been developed to address this shortcoming. The MDTB was created to provide a realistic test stand that effectively represents an operational environment and is able to bridge the chasm between typical university-scale test facilities and real-world applications.

The MDTB was constructed to collect calibrated, transitional data of both gear and bearing failures for commercial gearboxes and transmissions. The MDTB is instrumented with 52 sensors including 31 thermocouples, three internal temperature probes, seven single-axis accelerometers, a tri-axial accelerometer, a microphone, an acoustic emission sensor, an oil analysis sensor, a tachometer, two sets of torque and speed sensors, an infrared (IR) camera, and a borescope. Data are sampled using 16-channel, 16-bit DAQ boards. The sampling rate for the accelerometers is 20 kHz. Ten second snapshots of data are stored in a binary format to a PC. A detailed description of the PSU-ARL MDTB can be found in Reference12.

Data and Class Selection: As stated previously, the data to be used for the current effort should facilitate the development of prognostics by allowing features to be tracked during failure progression.

At the same time, if the data is to be realistically divided into classes, the system health status must be known by some "ground truth" capability. Recognizing this, researchers selected the transitional data from MDTB Run 14 for analysis. Run 14 employed a 3.33-ratio, single-reduction helical gearbox, and the test culminated with eight broken gearteeth. The ground truth for gearbox health is provided by the several borescope images that were obtained (Figure 1). Measurements were made and recorded periodically throughout the run using a variety of torque loads (an excerpt is shown in Figure 1) over the entire accelerated fault evolution. Accelerometer 3 (axial direction) was used for our analysis.

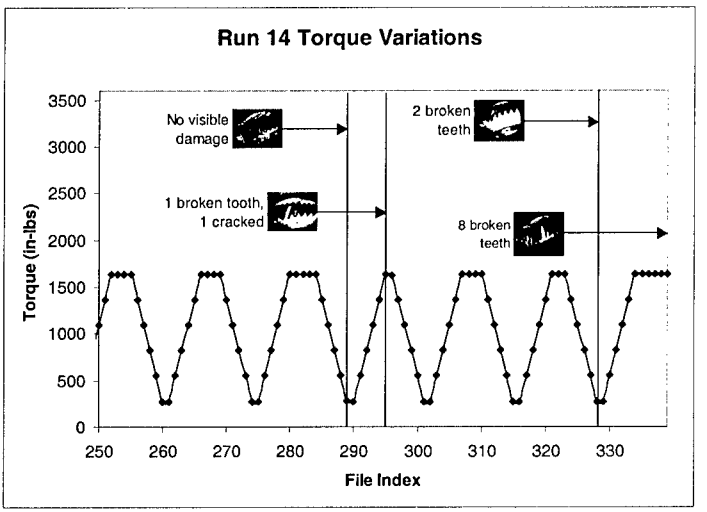

Figure 1: Run 14 Torque Variations, with Borescope Images Showing Gearbox Condition

The MDTB borescope data provides state points on the damage accumulation curve, but does not clearly identify the transition points (the ground truth occurs some time after the actual damage event). The data was divided into three classes: 1) no damage (snapshots 0-295), 2) 1-2 broken teeth (snapshots 296-328), 3) 2-8 broken teeth (snapshots 329-338). Although the faults began to occur before the borescope images were taken, basing the class boundaries solely on the borescope images is probably adequate, given the limited ground truth knowledge.

The features used in this analysis include: FM4, M8A, NA4*, INTR, INTSRC, and INTPK. FM4, M8A, and NA4* are common features that can be readily found in the literature [9,10]; the preprocessing is described in Reference 9. INTR (RMS), INTSRC (standard deviation of the rectified signal) and INTPK (spectrum peak magnitude of output shaft frequency) are features that were calculated on the ARL-developed interstitial signal [11].

Classification Results: Multiple cases were explored including: 1) using only the kurtosis to provide a baseline for the other results, 2) using six advanced features (FM4, NA4*, M8A, INTR, INTSRC, and INTPK) for the entire data set, and 3) using the six advanced features for only the high-torque data. Within each of these categories, a variety of classification and training methods were utilized. The advantage of using the advanced features should be apparent by the improvement in the error rate over the baseline.

Three classification methods are available in the SEPARAT toolbox: Parametric, Nonparametric, and Neural Network. The Nonparametric methods were used for the current investigation because there was not a good fit of the data with the built-in parametric technique (e.g., Gaussian classifier).

Baseline: Classification Using Kurtosis Alone: As can be seen in Table 1, using Kurtosis alone results in very high error rates (greater than 38%). Recall that the resubstitution training method provides a lower bound on the error and, thus, is the optimal result that can be achieved.

Table 1: Classification Errors when using Kurtosis Alone.

Classification Technique	Training/Testing Method	Error Rate
Fisher	Resubstitution	38.83%
MDE	Resubstitution	38.83%
MSE Linear	Resubstitution	66.86%
MSE Quadratic	Resubstitution	66.82%
Quadratic	Resubstitution	38.83%

Classification Using Advanced Features on the Entire Data Set: Results are provided in

Table 2 for simultaneous consideration of the six advanced features (classification using feature fusion). The results shown were obtained using resubstitution training, while the confusion matrix and error rate in Table 3 is for Leave-One-Out (LOO) training and the Minimum-Squared Error (MSE) classifier. We also transformed the six features into two space using a Fisher mapping technique, which yielded results slightly worse than in six space.

Given that resubstitution may be an optimistically biased method, an additional evaluation was performed on the most accurate classifier using LOO training to provide more realistic results. Results from this training method yielded a larger, yet more realistic error than was achieved using resubstitution (Table 3).

Table 2: Classification Errors using the Advanced Features

Classification Technique	Training/Testing Method	Error Rate
Fisher	Resubstitution	10.83%
MDE	Resubstitution	24.92%
MSE Linear	Resubstitution	16.79%
MSE Quadratic	Resubstitution	6.67%
Quadratic	Resubstitution	9.10%

Table 3: Confusion Matrix for MSE Quadratic Classifier using LOO Training Method and Advanced Features (Overall error = 18.63%)

	Class 1	Class 2	Class 3
Class 1	277	0	2
Class 2	0	28	5
Class 3	1	3	6

Feature Reduction: In many instances, multiple features will have a similar information basis and will not add significantly to the classification effectiveness. SEPARAT has the capability of ranking the features in terms of their contribution to class separability by exhaustively enumerating all possible combinations of features and noting which subsets maximize a selected criterion function (e.g., Fisher's criterion). The number of features to be included, r, is then selected by observing the plot and looking for a point where additional features do not seem to increase the criterion function. The vertical line is then dragged to a position between the r^{th} and $r^{th}+1$ points on the curve to choose the first r features within the ranking.

Figure 2 (left) is an output of SEPARAT that shows a ranking of the six features using Fisher's criterion. The figure shows that the use of more than the first three ranked features may not add significantly to the results and, thus, may not significantly impact the results. When using only the three highest ranked features (FM4, NA4*, M8A), the error increased from 6.67% (from fusion of the six features) to 13.58% (fusion of the three features). As the slope of the features in the feature ranking criterion plot approaches zero, the impact of these additional features on the results will diminish.

An additional evaluation was performed to map the six features into two space using the Fisher mapping technique. These results show that the six features can be transformed into two space with a resulting classification error of 10.83% using the LOO training method.

Figure 2 (right) shows the decision boundary results for this evaluation. This figure is useful for visualizing the distance between each classification point and the decision boundaries. In this case, significant overlap occurs between classes 2 and 3, which can be reasonably expected. Because the ground truth used to separate the classes is not

associated with a specific discrete degradation event, one should expect cross-over of classification in the neighborhood of the class boundary.

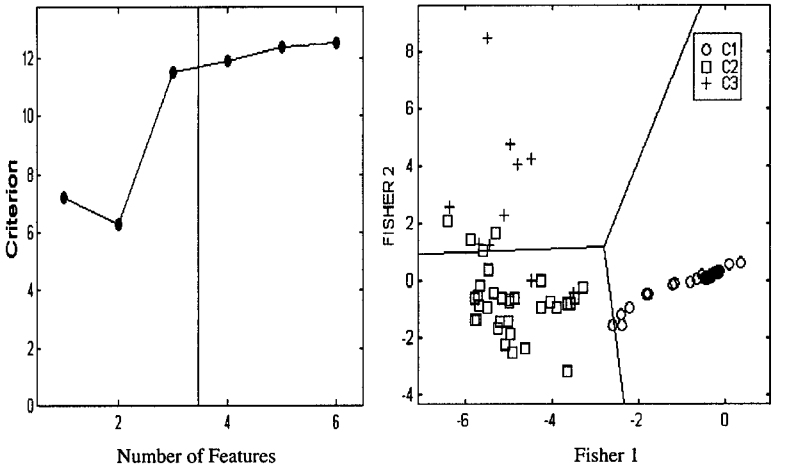

Figure 2: (Left) Feature Reduction Criterion Function. (Right) Decision Boundaries for Mapped Fisher in two-space.

Classification using Advanced Features and High Torque: The classification was repeated with a significantly truncated data set; using only the snapshots associated with high operational torque values. Various classifiers using the six advanced features were used, as well as classification using the Fisher mapping from six to two space. The results obtained using the truncated data set are provided below in Table 4.

Table 4 : Classification Error Rates for High Torque data

	Classification Technique	**Training/Testing Method**	**Error Rate**
Advanced Features	Fisher	Resubstitution	5.28%
	MDE	Resubstitution	21.39%
	MSE Linear	Resubstitution	13.06%
	MSE Quadratic	Resubstitution	11.94%
Fisher Mapping	Fisher	Resubstitution	4.17%

These results show very successful classification when only the high torque data is used for classification: less than 6% error using Fisher classifier and less than 5% error when the features are mapped into two space. These results demonstrate the advantage of identifying operational influences on the features, and then using this information to enhance the classification.

Conclusion: The results of the SEPARAT analysis show that classification performance can be improved by using some advanced diagnostic features and accounting for operational parameter (e.g., torque) changes during the failure progression. The effects of using a reduced set of features on the classification performance were evaluated using feature ranking and reduction methods as well as feature mappings from six space into a reduced feature space. These analysis results show the relative improvements that could potentially be gained by incorporating advanced features and mapping techniques into the classification scheme. The development of an optimal classification scheme would require a more critical selection of diagnostic features than those used herein.

Acknowledgements: This work was supported by the Office of Naval Research under the Accelerated Capabilities Initiative in Human Information Management through a subcontract by CHI Systems, Inc. (CHI-9803-002). The content of the information does not necessarily reflect the position or policy of the Government, and no official endorsement should be inferred.

References:

1. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, NY, 1973.
2. K. Fukunaga, *Statistical Pattern Recognition*, 2^{d} ed., Academic Press, San Diego, 1990.
3. J. P. Hoffbeck and D. A. Landgrebe, "Covariance Matrix Estimation and Classification With Limited Training Data," *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 18, No. 7, pp. 763-767, July 1996.
4. G. E. Golub and C. H. Van Loan, *Matrix Computations*, 2^{d} ed., The Johns Hopkins University Press, Baltimore, 1993.
5. K. Jain and M. D. Ramaswami, "Classifier Design with Parzen Windows," in *Pattern Recognition and Artificial Intelligence*, pp. 211-228, E. S. Gelsema and L. N. Kanal, eds. Elsevier Science Publishers B.V. (North-Holland), 1988.
6. K. Jain, R. C. Dubes, and C. C. Chen, "Bootstrap Techniques for Error Estimation," *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. PAMI-9, NO. 5, pp. 628-633, September 1987.
7. S. J. Raudys and A. K. Jain, "Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners," *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 13, No. 3, pp. 252-264, March 1991.
8. E. Parzen, "On Estimation of a Probability Density Function and Mode," *Ann. Math. Stat.*, 33, pp. 1065-1076, September 1962.
9. McClintic, K., et al, Residual and Difference Feature Analysis with Transitional Gearbox Data, 54th Meeting of the MFPT, Virginia, May 2000.
10. Lebold, M., et al, Review of Vibration Analysis Methods for Gearbox Diagnostics and Prognostics, 54th Meeting of the MFPT, Virginia, May 2000.
11. Maynard, K.P., *Interstitial Processing: The Application of Noise Processing to Gear Fault Detection*, Proceedings of International Conference on Condition Monitoring, Swansea, UK, 12-15 April 1999.
12. Byington, C.S., Kozlowski, J.D., "Transitional Data for Estimation of Gearbox Remaining Useful Life", 51st Meeting of the Society for Machinery Failure Prevention Technology (MFPT), April 1997.